Sequence-to-Sequence Data Augmentation for Dialogue Language Understanding

Yutai Hou, Yijia Liu, Wanxiang Che*, Ting Liu
Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
`{ythou, yjliu, car, tliu}@ir.hit.edu.cn`

Abstract

In this paper, we study the problem of data augmentation for language understanding in task-oriented dialogue system. In contrast to previous work which augments an utterance without considering its relation with other utterances, we propose a sequence-to-sequence generation based data augmentation framework that leverages one utterance's same semantic alternatives in the training data. A novel diversity rank is incorporated into the utterance representation to make the model produce diverse utterances and these diversely augmented utterances help to improve the language understanding module. Experimental results on the Airline Travel Information System dataset and a newly created semantic frame annotation on Stanford Multi-turn, Multi-domain Dialogue Dataset show that our framework achieves significant improvements of 6.38 and 10.04 F-scores respectively when only a training set of hundreds utterances is represented. Case studies also confirm that our method generates diverse utterances.

Title and Abstract in Chinese

对话语义理解的序列到序列数据增强

在本文中，我们研究了面向任务的对话系统中语言理解模块的数据增强问题。相比之前的工作在生成新语句时不考虑语句间关系，我们利用训练数据中与一个语句具有相同语义的其他句子，提出了基于序列到序列生成的数据增强框架。我们创新地将多样性等级结合到话语表示中以使模型产生多样化的语句数据，而这些多样化的新语句有助于改善语言理解模块。在航空旅行信息系统数据集以及一个新标注的斯坦福多轮多域对话数据集上的实验结果表明，当训练集仅包含数百句语料时，我们的框架在F值上分别实现了6.38和10.04的显着提升。案例研究也证实我们的方法能够产生多样化的话语。

1 Introduction

Language understanding (LU) is the initial and essential component in the task-oriented dialogue system pipeline (Young et al., 2013). One challenge in building robust LU is to handle myriad ways in which users express demands. This challenge becomes more serious when switching to a new domain whose large-scale labeled data is usually unreachable. Insufficiency in training data makes LU vulnerable to unseen utterances which are syntactically different but semantically related to the existing training data, and further harms the whole task-oriented dialogue system pipeline.

Data augmentation, which enlarges the size of training data in machine learning systems, is an effective solution to the data insufficiency problem. Success has been achieved with data augmentation on a wide range of problems including computer vision (Krizhevsky et al., 2012), speech recognition (Hannun et al., 2014), text classification (Zhang et al., 2015), and question answering (Fader et al., 2013). However, its application in the task-oriented dialogue system is less studied. Kurata et al. (2016a) presented the only work we know that tried to augment data for LU. In their paper, an encoder-decoder is

* Email corresponding.

learned to reconstruct the utterances in the training data. During the augmenting process, the encoder's output hidden states are randomly perturbed to yield different utterances.

The work of Kurata et al. (2016a) augments one single utterance by adding noise without considering its relation with other utterances. Besides theirs, there are also works which explicitly consider the paraphrasing relations between instances that share the same output. These works achieve improvements on tasks like text classification and question answering. Paraphrasing techniques including word-level substitution (Zhang et al., 2015; Wang and Yang, 2015), hand-crafted rules generation (Fader et al., 2013; Jia and Liang, 2016), and grammar-tree generation (Narayan et al., 2016) have been explored. Compared with these work, Kurata et al. (2016a) has the advantage of fully data-driven method and can easily switch to new domain without too much domain-specific knowledge, but doesn't make use of the relations between instances within the training data.

In this paper, we study the problem of data augmentation for LU and propose a novel data-driven framework that models relations between utterances of the same semantic frame in the training data. A sequence-to-sequence (seq2seq, Sutskever et al. 2014) model lies in the core of our framework which takes a delexicalised utterance and generates its lexical and syntactical alternatives. To further encourage diverse generation, we incorporate a novel *diversity rank* into the utterance representation. When training the seq2seq model, the diversity rank is also used to filter the over-alike pairs of alternatives. These approaches lead to diversely augmented data that significantly improves the LU performance in the domains that labeled data is scarce.

We conduct experiments on the Airline Travel Information System dataset (ATIS, Price 1990) along with a newly annotated layer of slot filling over the Stanford Multi-turn, Multi-domain Dialogue Dataset (Eric and Manning, 2017).[1] On the small proportion of ATIS which contains 129 utterances, our method outperforms the baseline by a 6.38 F-score on slot filling. On the medium proportion, this improvement is 2.87. Similar trends are witnessed on our LU annotation over Stanford dialogue dataset which the average improvement on three new domains is 10.04 on 100 utterances and 0.47 on 500 utterances.

The major contributions of this paper include:

- We propose a data augmentation framework for LU (§2) using the seq2seq model. A novel diversity rank (§3) is used to encourage our seq2seq model to generate diverse utterances both in the augmentation and training (§4).

- We conduct experiments on the ATIS and Stanford dialogue dataset (§5). Experimental results show our augmentation can effectively enlarge the training data and improve LU performance by a large margin when only a small size of training data is presented. Case studies also confirm that our method generates diverse utterances compared to the results from previous work.

We release our code at:
`https://github.com/AtmaHou/Seq2SeqDataAugmentationForLU.`

2 Overview of the Approach

Notion and Problem Description. In this paper, we study the data augmentation for language understanding (LU), which maps a natural language utterance into its semantic frames. We focus on *slot filling* and follow previous works (Pieraccini et al., 1992) by treating it as a sequence classification in which semantic class labels (*slot types*) are assigned to contiguous sequences of words indicating these sequences are corresponding *slot values*. In this paper, we use the bidirectional long short term memory (BiLSTM) for slot labeling (tagging) as previous works did (Mesnil et al., 2013; Yao et al., 2014; Kurata et al., 2016b).

We formalize the data augmentation for LU as given a natural language utterance $\mathbf{u}$ and its semantic frame $\mathbf{s}$, we generate a set of new utterances with corresponding semantic frames. During the augmenting process, we go through the whole training data $D = \{(\mathbf{u}_i, \mathbf{s}_i)\}_{i=1}^{N}$. For each training instance $(\mathbf{u}_i, \mathbf{s}_i)$,

---

[1]abbreviated as *Stanford dialogue dataset* henceforth.

show me the [closest]`<distance>`
[restaurant]`<poi_type>`

find me the **`<distance>` route to `<poi_type>`**
  (1.0) *give me the* `<distance>` *route to* `<poi_type>`
  (4.4) *i 'm desiring to eat at some* `<poi_type>`
    is there any in `<distance>`
  (5.0) *is there a* `<distance>` `<poi_type>`

| delexicalisation |

| ranking candidates by *diversity score* |

show me the `<distance>` `<poi_type>`

find me the **`<distance>` *route to* `<poi_type>`**
  #1: *is there a* `<distance>` `<poi_type>`
  #2: *i 'm desiring to eat at some* `<poi_type>`
    is there any in `<distance>`
  #3: *give me the* `<distance>` *route to* `<poi_type>`

| diverse ranks incorporation |

show me the `<distance>` `<poi_type>` #1
show me the `<distance>` `<poi_type>` #2
show me the `<distance>` `<poi_type>` #3

| filtering and generating "translation" pairs |

| seq2seq generation |

find me the `<distance>` *route to* `<poi_type>` #1
  → *is there a* `<distance>` `<poi_type>`
find me the `<distance>` *route to* `<poi_type>` #2
  → *i 'm desiring to eat at some* `<poi_type>`
    is there any in `<distance>`

where is the `<distance>` `<poi_type>`
can you find the `<distance>` `<poi_type>` *to me*
give me the address to the `<distance>` `<poi_type>`

| training model with filtered pairs |

| surface realisation |

| seq2seq model |

where is the nearest shopping mail
can you find the nearest rest stop to me
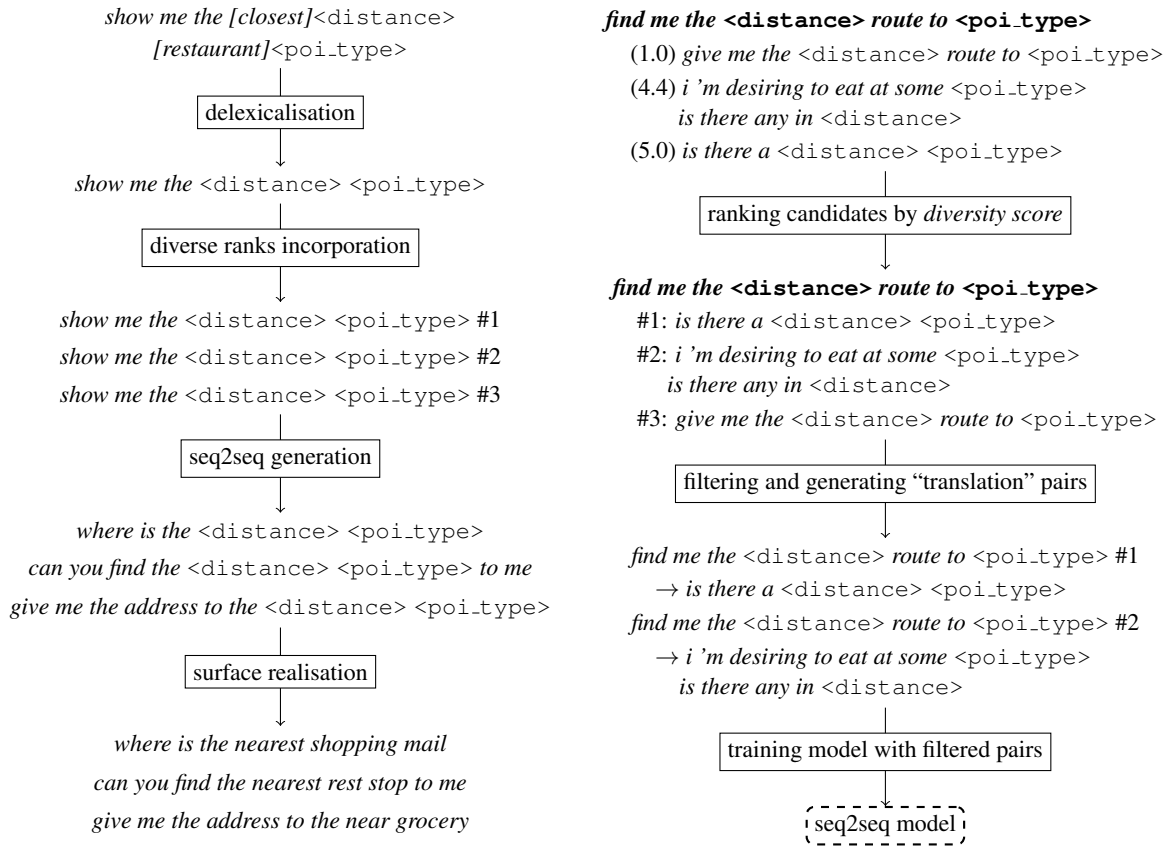*give me the address to the near grocery*

Figure 1: The workflow of our framework. The left part shows the augmenting process and the right part shows the training instance generation process for our seq2seq model. $\mathbf{u} \rightarrow \mathbf{u}'$ marks that $\mathbf{u}$ can be augmented into $\mathbf{u}'$.

we expand it to a set of instances $\{(\mathbf{u}_i^k, \mathbf{s}_i^k)\}_k$ and use the union of the expanded instances as new data to train the LU module.

In the training phase, we define the cluster of semantic frame $\mathbf{s}$ as $C_{\mathbf{s}} = \{(\mathbf{u}', \mathbf{s}') \mid (\mathbf{u}', \mathbf{s}') \in D \wedge \mathbf{s}' = \mathbf{s}\}$. For one utterance $\mathbf{u}$ and its semantic frame $\mathbf{s}$, each utterance $\mathbf{u}' \in C_{\mathbf{s}}/\{(\mathbf{u}, \mathbf{s})\}$ is considered as the alternative expression and augmentation of $\mathbf{u}$. We use $\mathbf{u} \rightarrow \mathbf{u}'$ to mark this relation.

To achieve the goal of generating variant utterances under the same semantic frames, We break down the problem into first converting the input utterance $\mathbf{u}$ into its delexicalised form $\mathbf{d}$, and then generating the delexicalised variances of $\mathbf{d}$ with a seq2seq model. Finally, surface realization is carried out to convert the delexicalised form into the raw utterance. The left part of Figure 1 shows the workflow of our augmenting process.

Delexicalisation. When given the raw utterance and its semantic frames associated with certain segments of the utterance, we can easily delexicalise the utterance by replacing the corresponding segments with the semantic frame label. For example, when given the 4th word in "*show me the closest restaurant*" as a `<distance>` slot type and 5th word as `<poi_type>` slot type, its delexicalized form "*show me the* `<distance>` `<poi_type>`" is straight-forward to achieve.

In the task-oriented dialogue system, slot values usually consist of various entity names and are very sparse. Delexicalisation reduces the size of vocabulary and makes the model focus more on generating variant ways of expressing demands. What's more, the semantic frames can be directly derived from the delexicalised generation and used for training the LU module.

Incorporating Diversity Ranks into Utterance Representations. Considering the example in the right part of Figure 1, "*is there a* `<distance>` `<poi_type>`" is more diverse than "*give me the* `<distance>` *route to* `<poi_type>`" when compared with "*find me the* `<distance>` *route to*

1236

`<poi_type>`". This example shows that for utterance **u** with semantic frame **s**, its alternatives expressions can have different ranks in diversity. To consider the ranking information, we compile the *diversity rank* as an additional information into the utterance representation. By setting it to a higher rank, we aim to generate input utterance's diverse augmentation, and by setting it to lower, a similar utterance should be generated. We will discuss the details of how to compute the ranks during training and how to decide the effective numbers of ranks during testing in Section 3.

Data Augmentation as Seq2Seq Generation. When given the delexicalised input utterance **d** and the specified diverse rank k, we use the standard seq2seq model to generate the alternative delexicalised utterance **d'**. In our seq2seq model, we append #k to the end of the input utterances and the model is formalized as

$$p(\mathbf{d}' \mid \mathbf{d}, k) = \prod_t p(d'_t \mid d_1, ..., d_n, \#k, d'_1, ..., d'_{t-1})$$

where n is the number of words for the input utterance **d**.

In this paper, we follow the seq2seq model for neural machine translation and use the *input-feeding* network in (Luong et al., 2015) with attention as our seq2seq model. During testing, we use beam search with beam size of 10 to yield more than one translation following Gimpel et al. (2013) and Vijayakumar et al. (2016).

To train the seq2seq model, our basic assumption is that if **d** and **d'** contain the same semantic frames, they can be generated from each other. Generally, we assume each pair of delexicalised utterances in the cluster $C_\mathbf{s}$ makes a pair of generation. However, it's nontrivial to assign *diverse ranks* to training data. What's more, to prevent the model from just producing produce lexical paraphrases (like "*show me*" to "*give me*"), we propose to also consider the diversities when generating training translations for the seq2seq model. We will talk about the details in Section 4.

Surface Realisation. Till now, we have achieved the lexically and syntactically different utterances in their delexicalized forms. We would like to bridge these utterances to their lexicalized forms and surface realisation is employed as the final step of our approach.

In this paper, the surface realisation is performed by replacing the slot type in the delexicalised form with its slot value. The mapping from slot type to its set of slot values (e.g. from `<poi_type>` to {*hospital*, *restaurant*}) is collected on the training data. Somehow, it's nontrivial to just do the replacement because one slot value doesn't fit its slot type in any context. Taking the utterance in Figure 1 for example, in the delexicalised utterance "*i 'm desiring to eat at some* `<poi_type>` *is there any in* `<distance>`", 'hospital' doesn't fit in the `<poi_type>` because 'hospital' isn't the place intended for a meal. To make the surface realisation more reasonable, we build the mapping with consideration of the context and use slot type along with its surrounding 5 words as the key in the mapping.

During surface realisation for an utterance, we first extract the slot type and its context. Then we use this to get all its slot values. If the slot type under certain context is not presented in the mapping, we use the one with the most similar context in the sense of *edit distance*. If more than one slot values present, we randomly pick a slot value.

3 *Diversity Ranks* in Utterance Representations

The major motivation of this paper is to encourage diverse generation. To accomplish this motivation, we propose a criterion named *diversity rank* to model the diversities. During augmenting the data, for an instance $(\mathbf{u}, \mathbf{s})$ we generate the delexicalised utterance at rank from 1 to $N_\mathbf{s}$, where $N_\mathbf{s}$ is a number governed by the semantic frame **s** and calculated as $||C_\mathbf{s}||/2$, which is the half size of the instances in D that have the semantic frame **s**.

During training the seq2seq model with diversity rank, for one instance $(\mathbf{u}, \mathbf{s})$, we first collect $C_\mathbf{s}$, then rank each instance $(\mathbf{u}', \mathbf{s}) \in C_\mathbf{s}/\{(\mathbf{u}, \mathbf{s})\}$ by its *diversity score* against **u**. In this paper, the diversity score of an utterance pair $(\mathbf{u}, \mathbf{u}')$ is calculated by both considering the *edit distance* and a *length difference penalty* (LDP) as:

$$\text{SCORE}(\mathbf{u}, \mathbf{u}') = \text{EDITDISTANCE}(\mathbf{u}, \mathbf{u}') \times \text{LDP}(\mathbf{u}, \mathbf{u}') \tag{1}$$

	Navigation	Scheduling	Weather
# of training utterances	500	500	500
# of devel. utterances	321	201	262
# of test utterances	337	212	271
Kappa	0.68	0.92	0.90
Agreement	85.05	90.75	95.99

Table 1: Statistics for our annotation.

where LDP is defined as $\text{LDP}(\mathbf{u}, \mathbf{u}') = e^{-\frac{|||\mathbf{u}||-||\mathbf{u}'|||}{||\mathbf{u}||}}$. After obtaining the ranks over the utterances $\mathbf{u}'$, we directly incorporate the rank value as an additional last token for the seq2seq model.

We note that using the LDP reduces the impact of differences in length and makes the score paying more attention to the lexical and syntactical difference. For example, the first block of right part of Figure 1 shows the diversity scores of three different utterances. Although the utterance "*i 'm desiring to eat at some* `<poi_type>` *is there any in* `<distance>`" presents larger *edit distance* (12 in this case) than that of "*is there a* `<distance>` `<poi_type>`" (5 in this case), the final score is penalized to 4.4 because the length difference.

In our method, the diversity rank can be treated as an utterance-independent controller for the diversity of target generation.

4 Filtering the Alike Instances

To learn the seq2seq model, it's straight-forward to use each pair of utterances in $C_{\mathbf{s}}$ as training data for the model. However, the goal of our paper is to generate diverse augmented data and the usefulness of less diverse pair (like *give me the* `<distance>` *route to* `<poi_type>` and *find me the* `<distance>` *route to* `<poi_type>` in Figure 1) is arguable.

In this paper, we propose to filter the less diverse pairs when training the seq2seq model. Again, we make use of the ranks derived by the diversity scores and for an utterance $\mathbf{u}$ only the most diverse half of the translations $\mathbf{u} \rightarrow \mathbf{u}'$ are used to train the seq2seq model and the training data can be formalized as

$$D_{\text{seq2seq}} = \bigcup_{(\mathbf{u},\mathbf{s})\in D} \{\mathbf{u}, \text{RANK}(\mathbf{u}, \mathbf{u}') \rightarrow \mathbf{u}' \mid \mathbf{u}' \in C_{\mathbf{s}}, \text{RANK}(\mathbf{u}, \mathbf{u}') \geq ||C_{\mathbf{s}}||/2\}$$

After filtering the less diverse pairs, we use D_{seq2seq} to train the seq2seq model.

In this section, we revisit the role of our diversity ranks in the learning perspective. Since we consider the utterance in cluster $C_{\mathbf{s}}$ as translation to each other, without the RANK value, one utterance can simultaneously translate to different utterances in the training data. It increases the ambiguities in learning the seq2seq model and even makes it intractable. With the RANK value, such ambiguities are resolved because each pair of the training data is expanded with a unique value.

5 Experiments

5.1 Settings

Dataset. In this paper, we conduct our experiments on the ATIS dataset which is extensively used for LU (Mesnil et al., 2013; Mesnil et al., 2015; Chen et al., 2016a). The ATIS dataset contains 4978 training utterances from Class A training data in the ATIS-2 and ATIS-3 corpus, while the test contains 893 utterances from the ATIS-3 Nov93 and Dec94 datasets. The size of the training data is relatively large for LU in a single domain. To simulate the data insufficient situations, we follow Chen et al. (2016a), and also evaluate our model on two small proportions of the training data which is *small* (1/40 of the original training set with 129 instances) proportion and *medium* (1/10 of the original training set with 515 instances). In all the experiments, a development set of 500 instances is used.

To test our model on new domains beyond ATIS, we also create a new LU annotation over the Stanford dialogue dataset (Eric and Manning, 2017). We use the same data split as Eric and Manning (2017) and

Model	small	medium	full
	129	515	4,478
Baseline	67.33**	85.85**	94.93*
Ours	73.71	88.72	94.82
Re-implementation of Kurata et al. (2016a)	67.93**	87.34**	94.61**
Model-1 Additive (Kurata et al., 2016a)	-	-	95.08
K-SAN syntax (Chen et al., 2016a)	74.35	88.40	95.00
Model-III (Zhai et al., 2017)	-	-	95.86

Table 2: The results on the ATIS dataset. The first block shows the results from our implementation and the second block is drawn from the papers of previous works. Here we use * to indicate that the difference between the model and Ours is statistically significant under t-test (** for p-value threshold as 0.05 and * for threshold as 0.1) .

annotate the full test sets for the three domains (*navigation*, *scheduling*, and *weather*) along with a small training set of 500 utterances. The Stanford dialogue dataset provides semantic frames (*slot*) for each utterance but doesn't associate the semantic class of the slot with corresponding segment in the utterance. Our annotation focus on assigning the slot to its corresponding segment. During the annotation, each dialogue was processed by two annotators. Data statistics, Kappa value (Snow et al., 2008), and inner annotator agreement measured by F-score on the three domains are shown in Table 1.

Evaluation. We evaluate our data augmentation's effect on LU with F-score. `conlleval` is used in the same way with previous works (Mesnil et al., 2013; Mesnil et al., 2015; Chen et al., 2016a).

Implementation. We use OpenNMT (Klein et al., 2017) as the implementation of our seq2seq model. We set the number of layers in LSTM as 2 and the size of hidden states as 500. Utterances that are longer than 50 are truncated. We adopt the same training setting as Luong et al. (2015) and use Adam (Kingma and Ba, 2014) to train the seq2seq model. Learning rate is halved when perplexity on the development set doesn't decrease. During generation, we replace the model-yielded unknown token (*unk*) with the source word that has the highest attention score.

For the slot tagging model, we set both the dimension for word embedding and the size of hidden state to 100. We also vary dropout rate in {0, 0.1, 0.2} considering its regularization power on small size of data. The batch size is set to 16 in all the experiments. Best hyperparameter settings are determined on the development set. GloVe embedding (Pennington et al., 2014) is used to initialize the word embedding in the model. Adam with the suggested settings in Kingma and Ba (2014) is used to train the parameters.

Reimers and Gurevych (2017) pointed out that neural network training is nondeterministic and depends on the seed for the random number generator. We witness dramatic changes of the slot tagging performance using different random seeds. To control for this effect, we take their suggestions and report the average of 5 differently-seeded runs.

5.2 Results on ATIS

Table 2 shows the slot tagging results on the ATIS dataset. Our baseline model is the vanilla BiLSTM slot tagger and our augmented slot tagger use the same architecture but is trained with the augmented data generated by our method. Compared with the vanilla tagger baseline, our augmentation method significantly improves the LU performance by a 6.38 F-score on the *small* proportion and a 2.02 F-score on the *medium* proportion. The improvements show the effectiveness of our augmentation method in the data-insufficient scenario. On the full data, our augmentation slightly lags the baseline. We address this to the fact that full ATIS is large enough for LU on a single domain and our augmentation introduce some noise.

To compare with the previous augmentation work from Kurata et al. (2016a), we re-implemented their *model-1 additive* model using the suggested settings in their paper. The results on the *small*, *medium*, and *full* proportions are shown in the third row of Table 2. On all the proportions, our augmentation

# utterances	Model	Navigation	Scheduling	Weather
100	Baseline	59.93	68.29	82.43
	Ours	72.91	77.30	90.55
500	Baseline	78.99	86.05	93.68
	Ours	78.46	87.67	94.01

Table 3: The results on Stanford dialogue dataset.

Model	F-score	# new	max. ED
Ours	88.72	301	3.18
- seq2seq generation	-0.84**	0	0
- diversity ranks	-0.40*	163	2.42
- filtering	-0.38	870	2.86

Table 4: The result of the ablation test. *# new* marks the number of newly generated delexicalised utterances. *max. ED* marks the averaged maximum edit distances. Here we use * to indicate that the result is statistically significant under t-test (** for p-value threshold as 0.05 and * for threshold as 0.1) By removing the seq2seq generation from our method, no delexicalised utterance will be generated so the *max. ED* cell is 0.

method outperforms theirs and the differences are significant on *small* and *medium*. Since their model relies on learning a seq2seq model to reconstruct the input utterances, it's usually difficult to train a reasonable model on very small data due to sparsity. Our method mitigates this by both generating on the delexicalised utterances and learning the generation model from pairs of utterances that share same semantic frame which enlarge the size of data for us to train the model. We also compare our model with the syntax version of K-SAN (Chen et al., 2016a) without joint training from intent annotation. We see that our augmented tagger lags their syntax-parsing-enhanced model by a 0.64 F-score on small proportion and outperforms theirs by a 0.32 F-score on medium proportion. But considering the training data is sampled with different random seeds between our work and theirs, these results are not directly comparable. At last, we show the (Zhai et al., 2017) as state-of-art results on ATIS dataset, which views slot filling task as sequence chunking problem. As we focus data augmentation for sequence labeling task rather than chunking, this result is not directly comparable to ours. Besides, K-SAN (Chen et al., 2016a) and (Zhai et al., 2017) are not data augmentation methods, we included their results to show that our augmentation method is reasonably good The basic trend shows that our augmentation can be used as an alternative to the LU model leveraging rich syntactic information.

5.3 Results on Stanford Dialogue Dataset

The results for Stanford dialogue dataset are shown in Table 3. Similar trend as the ATIS experiments is witnessed in which the augmentation improves the LU performance. The average improvement on the training data with 100 utterances is 10.04, and the number is 0.47 for that with 500 utterances. Considering that only fewer than 350 utterances present in the test set in all these domains, these improvements are reasonable. Besides, similar to the ATIS results, the margin of improvements is larger for the smaller training set.

An advantage of our method is that it's purely data-driven. Only a mapping from slot type context to slot values is required and it can be constructed from the training data. It's easy for our method to switch to new domains and our results on the Stanford dialogue dataset confirms this.

5.4 Analysis

Ablation. To get further understanding of each component in our method, we conduct ablation on the *medium* proportion, Each of the three parts of our method is removed respectively, including the *seq2seq generation*, *diversity ranks*, and *filtering*. In addition to evaluate the model's performance with F-score, we also examine the augmented data by the number of newly generated delexicalised utterances and the
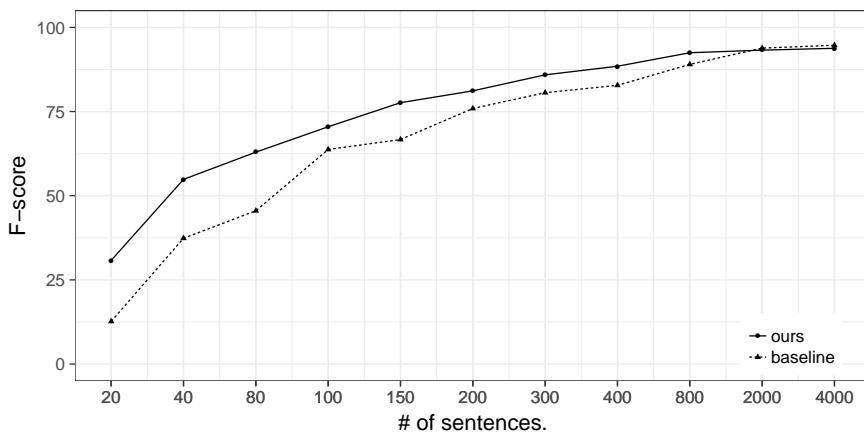
Figure 2: Our method's performances on the ATIS training data of different sizes.

maximum edit distances against the rest of instances.[2] The results are shown in Table 4.

For our method without *seq2seq generation*, we only conduct surface realisation on the delexicalised utterance and a 0.84 F-score drop is witnessed. Since surface realisation only substitutes slot type with different slot values without changing the utterances syntactically, this ablation shows it's more beneficial to generate syntactic alternatives using our seq2seq model.

For our method without *diversity ranks*, we remove diversity ranks from the utterance representation and this lead a drop of 0.40 F-score. We address the drop of performance to the fact that removing either these components will lead to less diverse generation. The second and third column in Table 4 confirm this by showing less newly and diversely generated delexicalised utterances.

If we don't filter the alike instances when training the seq2seq model, the drop of performance is a 0.65 F-score. However, larger number of new utterances with smaller edit distances are yielded which indicates that more noise is introduced when the training data of the seq2seq model is not properly filtered.

This ablation also shows correlation between the maximum edit distance and the final F-score, which indicates generating diverse augmentation helps the performance.

Effect of Training Data Size. The results on ATIS and Stanford dialogue dataset witness the trend that smaller training data benefits more from our augmentation method. A natural question that arises is what's boundary of our augmentation in the sense of improving the baseline. In this section, we study this by varying training data size on the ATIS data. Figure 2 shows the results. For the ATIS data, improvements can be achieved in all our settings with training size smaller than one thousand. These results indicate that our augmentation is applicable when we only access to a LU training data of hundreds instances.

Case Study. In this paragraph, we perform case study on our method to verify its capability of generating diversely augmented data. Table 5 shows two cases of our augmentation. Each case includes the original sentence and its delexicalised form (in *italic* font), the diversity rank (starts with # mark), the training utterance under this rank, our augmentation along with surface realization, and the augmentation produced by Kurata et al. (2016a).

By comparing our augmentation with the delexicalised form of source utterance, two observations can be drawn: 1) our method yields syntactically different alternatives meanwhile keeps the original semantic frame as the source utterance; 2) the lengths of the generated utterances are in the same scale with the source utterance thanks to the effect of length penalty in Equation 1.

By comparing our augmentation with the target training utterance under the same rank, our seq2seq model yields different utterance instead of repeating the training utterance. We address this diversity

---

[2]This number is normalized by the total number of utterances.

		show me all flights from atlanta to washington with prices
		(delex.) *show me all flights from* `<from_city>` *to* `<to_city>` *with prices*
#1	train	let 's look at `<from_city>` to `<to_city>` again
	ours	what are all the flights between `<from_city>` and `<to_city>`
		(realized) what are all the flights between indianapolis and tampa
#100	train	list types of aircraft that fly between `<from_city>` and `<to_city>`
	ours	i 'm looking for a flight from `<from_city>` to `<to_city>`
		(realized) i 'm looking for a flight from milwaukee to los angeles
	Kurata16	show me all flights from [atlanta]`<from_city>` to [washington]`<to_city>` with airports

		is there a flight between san francisco and boston with a stopover at dallas fort worth
		(delex.) *is there a flight between* `<from_city>` *and* `<to_city>` *with a stopover at* `<stop_city>`
#1	train	which airlines fly from `<from_city>` to `<to_city>` and have a stopover in `<stop_city>`
	ours	is there a flight from `<from_city>` to`<to_city>` with a stop in `<stop_city>`
		(realized) is there a flight from washington to miami with a stop in dallas fort worth
#30	train	do you have any airlines that would stop at `<stop_city>` on the way from `<from_city>` to `<to_city>`
	ours	i 'd like to fly from `<from_city>` to `<to_city>` with a stop in `<stop_city>`
		(realized) i 'd like to fly from memphis to boston with a stop in minneapolis
	Kurata16	is there a flight between [san francisco]`<from_city>` and [boston]`<to_city>` with a stopover at [dallas fort worth]`<to_city>`

Table 5: Case study of our augmented data against the training data and the results of Kurata et al. (2016a) (marked as Kurata16). *train* marks the target utterance in the training data. **(delex.)** marks the delexicalised form of the input utterance. **(realized)** marks the utterance after surface realisation.

to the fact that our diversity rank has some universal effect on modeling the diversity degree across different instances. When contrasting to the augmentation of Kurata et al. (2016a), our method clearly shows diverse augmentation against the source utterance while theirs are basically repeating the source utterances. In the sense of generating diverse alternatives for expressing the same semantics, our method has the advantage.

6 Related work

Data augmentation is an effective way of improving the model's performance and it has been extensively explored on the computer vision community. Single transformation approaches like randomly copying, flipping, and changing the intensity of RGB are the common practice in the top-performed vision systems (Krizhevsky et al., 2012). Beyond these classic approaches, adding noise to the image, randomly interpolating a pair of images (Zhang et al., 2018) are also proposed in previous works. However, these signal transformation approaches are not directly applicable to language because order of words in language may form rigorous syntactic and semantic meaning (Zhang et al., 2015). Therefore, the best way of data augmentation in language usually involves generating the alternative expressions.

Paraphrasing is the most studied techniques in natural language processing for generating alternative expressions (Barzilay and McKeown, 2001; Bannard and Callison-Burch, 2005; Callison-Burch, 2008). However, generic paraphrasing technique has been reported not helpful for specific problem (Narayan et al., 2016). Most of the successful work that applying paraphrasing for data augmentation requires special tailored paraphrasing techniques. For example, Wang and Yang (2015) performed word-level paraphrasing to extend their corpus on twitter that contains annoying behaviors. Fader et al. (2013) derived question templates from seed paraphrases and bootstrap the templates to achieve the enlarged open-domain QA dataset. Narayan et al. (2016) constructed latent variable PCFG for questions and augment the training data by sampling from the grammar. All these works assume the same output (i.e.

class in text classification, answer in question answering) for input paraphrases. Our method resembles theirs in the assumption for input paraphrases, but differs on using the seq2seq generation which is purely data-driven and doesn't rely on special tailored domain knowledge. Besides these methods, works that introduce errors to language understanding have also been proposed (Schatzmann et al., 2007b; Sagae et al., 2012).

Language understanding, as an important component in the task-oriented dialogue system pipeline, has drawn a lot of research attention in recent year, especially when enhanced by the rich representation power of the neural network, like recurrent neural network, LSTM (Yao et al., 2013; Yao et al., 2014; Mesnil et al., 2013; Mesnil et al., 2015) and memory network (Chen et al., 2016b). Rich linguistic features (Chen et al., 2016a) and representation in broader scope on sentence-level (Kurata et al., 2016c) and dialogue history-level (Chen et al., 2016b) have also been studied. Our augmentation method is orthogonal to these works and it's hopeful to achieve more improvements with their works.

Dialogue management is also a key component of task-oriented dialogue system, which mainly focuses on dialogue policy. However, optimal dialogue policy is hard to obtain from a static corpus due to the vast space of conversation process. A solution is to transform the static corpus into user simulator (Kreyssig et al., 2018), and most user simulators work on user semantics level. (Eckert et al., ; Schatzmann et al., 2007a; Asri et al., 2016; Scheffler and Young, 2000; Scheffler and Young, 2001; Pietquin and Dutoit, 2006; Georgila et al., 2005; Cuayáhuitl et al., 2005). Recent work starts to generate user utterance directly to reduce data annotation(Kreyssig et al., 2018).

In recent years, Generative Adversarial Network (GAN, Goodfellow et al. 2014) draws a lot of research attention. Its ability of generating adversarial examples is attractive for data augmentation. However, it hasn't been tried in data augmentation beyond computer vision (Antoniou et al., 2018). How to apply GAN to language understanding is still an open question.

7 Conclusion

In this paper, we study the problem of data augmentation for LU. We propose a data-driven framework to augment training data. In our framework, one utterance's alternative expressions of the same semantic are leveraged to train seq2seq model. We also propose a novel diversity rank to encourage diverse generation and filter alike instances. In the experiments, our model achieves significant improvements of 6.38 and 10.04 F-scores respectively when only a training set of hundreds utterances is represented. Careful case study also shows the capability of our framework to generate diverse alternative expressions.

Acknowledgements

References

Anthreas Antoniou, Amos Storkey, and Harrison Edwards. 2018. Data augmentation generative adversarial networks.

Layla El Asri, Jing He, and Kaheer Suleman. 2016. A sequence-to-sequence model for user simulation in spoken dialogue systems. *arXiv preprint arXiv:1607.00070*.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proc. of ACL*.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proc. of ACL*.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proc. of EMNLP*.

Yun-Nung Chen, Dilek Hakanni-Tür, Gokhan Tur, Asli Celikyilmaz, Jianfeng Guo, and Li Deng. 2016a. Syntax or semantics? knowledge-guided joint semantic frame parsing. In *SLT*, pages 348–355.

Yun-Nung Vivian Chen, Dilek Hakkani-Tür, Gokhan Tur, Jianfeng Gao, and Li Deng. 2016b. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In *INTERSPEECH*.

Heriberto Cuayáhuitl, Steve Renals, Oliver Lemon, and Hiroshi Shimodaira. 2005. Human-computer dialogue simulation using hidden markov models. In *ASRU*, pages 290–295. IEEE.

W. Eckert, E. Levin, and R. Pieraccini.

Mihail Eric and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. *arXiv preprint arXiv:1705.05414*.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In *Proc. of ACL*.

Kallirroi Georgila, James Henderson, and Oliver Lemon. 2005. Learning user simulations for information state update dialogue systems. In *Ninth European Conference on Speech Communication and Technology*.

Kevin Gimpel, Dhruv Batra, Chris Dyer, and Gregory Shakhnarovich. 2013. A systematic exploration of diversity in machine translation. In *Proc. of EMNLP*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition. *CoRR*, abs/1412.5567.

Robin Jia and Percy Liang. 2016. Data recombination for neural semantic parsing. In *Proc. of ACL*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proc. of ACL 2017, System Demonstrations*.

Florian Kreyssig, Inigo Casanueva, Pawel Budzianowski, and Milica Gasic. 2018. Neural user simulation for corpus-based policy optimisation for spoken dialogue systems. *arXiv preprint arXiv:1805.06966*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Gakuto Kurata, Bing Xiang, and Bowen Zhou. 2016a. Labeled data generation with encoder-decoder LSTM for semantic slot filling. In *INTERSPEECH 2016*, pages 725–729.

Gakuto Kurata, Bing Xiang, Bowen Zhou, and Mo Yu. 2016b. Leveraging sentence-level information with encoder lstm for semantic slot filling. In *Proc. of EMNLP*.

Gakuto Kurata, Bing Xiang, Bowen Zhou, and Mo Yu. 2016c. Leveraging sentence-level information with encoder lstm for semantic slot filling. *arXiv preprint arXiv:1601.01530*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proc. of EMNLP*.

Grégoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio. 2013. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *INTERSPEECH 2013*.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Z. Hakkani-Tür, Xiaodong He, Larry P. Heck, Gökhan Tür, Dong Yu, and Geoffrey Zweig. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM TASLP*, 23(3):530–539.

Shashi Narayan, Siva Reddy, and Shay B. Cohen. 2016. Paraphrase generation from latent-variable pcfgs for semantic parsing. In *Proc. of INLG*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

R. Pieraccini, E. Tzoukermann, Z. Gorelov, J. L. Gauvain, E. Levin, C. H. Lee, and J. G. Wilpon. 1992. A speech understanding system based on statistical representation of semantics. In *Proc. of ICASSP*, Mar.

Olivier Pietquin and Thierry Dutoit. 2006. A probabilistic framework for dialog simulation and optimal strategy learning. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):589–599.

P. J. Price. 1990. Evaluation of spoken language systems: The atis domain. In *Proc. of the Workshop on Speech and Natural Language*, HLT '90.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proc. of EMNLP*.

Kenji Sagae, Maider Lehr, E Prud'hommeaux, Puyang Xu, Nathan Glenn, Damianos Karakos, Sanjeev Khudanpur, Brian Roark, Murat Saraclar, Izhak Shafran, et al. 2012. Hallucinated n-best lists for discriminative language modeling. In *ICASSP*, pages 5001–5004. IEEE.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007a. Agenda-based user simulation for bootstrapping a pomdp dialogue system. In *NAACL*, pages 149–152. Association for Computational Linguistics.

Jost Schatzmann, Blaise Thomson, and Steve Young. 2007b. Error simulation for training statistical dialogue systems. In *ASRU*, pages 526–531. IEEE.

Konrad Scheffler and Steve Young. 2000. Probabilistic simulation of human-machine dialogues. In *ICASSP*, volume 2, pages II1217–II1220. IEEE.

Konrad Scheffler and Steve Young. 2001. Corpus-based dialogue simulation for automatic strategy learning and evaluation. In *NAACL*, pages 64–70.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proc. of EMNLP*, pages 254–263.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *CoRR*, abs/1610.02424.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In *Proc. of EMNLP*, pages 2557–2563.

Kaisheng Yao, Geoffrey Zweig, Mei-Yuh Hwang, Yangyang Shi, and Dong Yu. 2013. Recurrent neural networks for language understanding. In *Interspeech*, pages 2524–2528.

K. Yao, B. Peng, Y. Zhang, D. Yu, G. Zweig, and Y. Shi. 2014. Spoken language understanding using long short-term memory neural networks. In *2014 IEEE SLT*, pages 189–194, Dec.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proc. of the IEEE*, 101(5):1160–1179.

Feifei Zhai, Saloni Potdar, Bing Xiang, and Bowen Zhou. 2017. Neural models for sequence chunking. In *AAAI*, pages 3365–3371.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.